SCHOOL OF COMPUTER SCIENCE
Faculty of Engineering and Information Technology
University of Technology Sydney

Hybrid Words Representation for the classification of low quality text

by

Usman Naseem

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Analytics (Research)

Sydney, Australia

2020

Certificate of Authorship/Originality

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as a part of the requirements for other degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis.

This research is supported by the Australian Government Research Training Program (RTP).

Production Note:
Signature removed prior to publication.

Acknowledgements

Acknowledgment goes here Foremost, I would like to express my sincere gratitude to my supervisor; Professor Longbing Cao for the continuous support of my master research degree, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor during my study.

I also would like to appreciate my co-supervisor A\Professor Kaska Musial-Gabrys for providing me with continuous support throughout my study and research. Without her professional guidance and persistent help, this thesis would not have been possible. Each of my supervisors helped, supported, and guided me through my research at UTS exceptionally and unforgettably. I shall always remain thankful to my supervisors.

I thank my fellow lab-mates in Advanced Analytics Institute (AAi): for the stimulating discussions, guidance, support and for all the fun we have had in the last two years. I am thankful to the office staff at the School of computer science, UTS, in particular, Margot, Janet, Teraesa and Reshma. All team members I met within the School, provided a conducive research environment. I am also thankful to the staff at the Graduate Research School at UTS, who always remained welcoming of my questions and queries, resolving my problems in a timely manner.

Last but not least, I would like to thank my family for their unconditional support, both financially and emotionally throughout the whole master studying.

Dedication

My thesis is dedicated to the people who have supported my goals, inspired me and challenged me academically to make it to this day.

Reflecting on my path that led me to this day, after spending more than eight years in the industry and having a very comfortable life with a good job, it was very hard decision for me to come back to student life with limited income. But I decided to take this challenge and thought may be I can do it. I pushed myself hard throughout this period. Today, I submit my master by research degree at the UTS and change my earlier 'maybe' into 'yes'.

For this day today, first and foremost credit goes to my parents. My father taught me to make decisions and always encouraged and supported my bold decision. In all bad times, he encouraged and supported me. I wish I could share this moment with my mother (late). Whatever I am today is only because of her prayers. Thanks to my siblings, Salman and Haadia for always remaining a great source of support and encouragement. My grandfather (late) who always prayed for my success. I still remember that I used to call my grandfather and asked him to pray for me whenever I used to feel down.

Without the support of my wife Huda, I would have never been able to complete my master degree. Aashir and Mahd, I owe you a lot. I could not give more time to you in your childhood due to coming back to home late nights during the weekdays and spending most of my weekends in the library doing my research - you were the real owners of this time. Huda, I am eternally thankful to you for taking up all the responsibilities and letting me focus on my research. Thank you. Thank you to my father-in-law and mother-in-law and as well as siblings of my wife who offered well wishes for my studies and prayed for my success

Contents

List of Figures

List of Tables

List of Publications

Paper(s) Accepted & Published

1. **Naseem, U**., Musial, K. (2019, September). Dice: deep intelligent contextual embedding for Twitter sentiment analysis. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 953-958). IEEE.
2. **Naseem, U.**, Khan, S. K., Razzak, I., Hameed, I. A. (2019, December). Hybrid Words Representation for Airlines Sentiment Analysis. In Australasian Joint Conference on Artificial Intelligence (pp. 381-392). Springer, Cham.
3. **Naseem, U.**, Razzak, I., Hameed, I. A. Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. Australian Journal of Intelligent Information Processing Systems, 69.
4. **Naseem U**., Khan S.K., Farasat M., Ali F: Abusive language detection: A Comprehensive Review: Indian Journal of Science and Technology., IJST,2019, **ACCEPTED**

Paper(s) to be Submitted & Under review

5. A Comparative Analysis of Pre-processing Techniques on Twitter Abusive language and Hate Speech detection.
6. A Recommended combination of pre-processing techniques to improve quality of text data.
7. A Comprehensive Survey on Word Representation Methods for Text Classification.
8. Hybrid Words Representation for text classification (Extended version of accepted paper)

Abbreviation

NLP - Natural Language Processing

HCI - Human-Computer Interaction

NLU- Natural Language Understanding

SA- Sentiment Analysis

TSA- Twitter Sentiment Analysis

OM- Opinion Mining

SMS- Short Message Service

RT- Retweet

DICE- Deep Intelligent Contextual Embedding

OOV- Out Of Vocabulary

SVM- Support Vector Machine

NB- Naive Bayes

LR- Logistic Regression

DT- Decision Tree

RF- Random Forest

DL- Deep Learning

BoW- Bag Of Word

TF- Term Frequency

TF-IDF- Term Frequency Inverse Document Frequency

CBOW- Continous Bag Of Words

GloVe- Global Vectors

CoVe- Context Vectors

LM- Language Model

BiLM- Bidirectional Language Model

ELMo- Embedding from language model

RNN- Recurrent Neural Network

LSTM- Long Short Term Memory

GRU- Gated Recurrent unit

BiLSTM- Bidirectional Long Short Term Memory

CNN- Convolutional Neural Network

UNK- Unique

POS- Part Of Speech

ReLu- Rectified Linear Unit

DCNN- Deep Convolutional Neural Network

HyRank- Hybrid Ranking

Re*- Refine Embedding

IWV- Improved Word Vectors

ABSTRACT

Hybrid Words Representation

for the classification of

low quality text

by

Usman Naseem

Language enables humans to communicate with others. For instance, we talk, give our opinions and suggestions all using natural language; to be more precise, we use words while communicating with others. However, in today's world, we wish to communicate with computers, just like humans. It is not an easy task because human communicate in an unstructured and informal way, whereas computers need structured and clean data. So it is essential for computers to understand and classify text accurately for proper human-computer interactions. For classifying a text, the first question we must address is how to improve the low-quality text. The next immediate challenge is to have the best representation so that text can be classified accurately. The way text is organized reflects polysemy, semantic and syntactical coupling relationships which are embedded in its contents. The effective capturing of such content relationships is thereby crucial for a better understanding of text representations. This is especially challenging in the environments where the text messages are short, informal and noisy, and involves natural language ambiguities. The existing sentiment classification methods are mainly for document and clean textual data which can not capture relationship, different attributes and characteristics within tweet messages.

Social media analysis, especially the analysis of tweet messages on Twitter has become increasingly relevant since the significant portion of data is ubiquitous in nature. The social media-based short text is valuable for many good reasons, ex-

plored increasingly in text analysis, social media analysis and recommendation. In the same time, there is a number of challenges that need to be addressed in this space. One of the main issues is that the traditional word embeddings are unable to capture polysemy (assigns the same representation of a word irrespective of its context and meaning) and out of vocabulary words (assigns a random representation). Furthermore, traditional word embeddings fail to capture sentiment information of words which results in similar word vector representations having the opposite polarities. Thus, ignoring polysemy within the context and sentiment polarity of words in a tweet reduces the performance for tweets classification.

In order to address the above-mentioned research challenges and limitations associated with word-level representations, this thesis focuses on improving the representation of low-quality text by improving the unstructured and informal nature of tweets to utilize the information thoroughly and manages the natural language ambiguities to build a more robust sentiment classification model. As compared to previous studies, the proposed models can deal with the ubiquitous nature of the short text, polysemy, semantic and syntactical relationships within a content, thereby addressing the natural language ambiguity problems.

Chapter 4 presents the effects of pre-processing techniques using two different word representation models with the machine and deep learning classifiers. Then, we present our recommended combination (approach) of different pre-processing techniques which improves the low quality, by performing sentiment-aware tokenization, correction of spelling mistakes, word segmentation and other techniques to utilize most of the information hidden in unstructured text. The experimental result shows that the proposed combination performs well as compared to other combinations.

Chapter 5 presents the hybrid words representation. In this chapter, we proposed our Deep Intelligent Contextual Embedding for Twitter sentiment analysis. Proposed model addresses the natural language ambiguities and is devised to capture polysemy in context, semantics, syntax and sentiment knowledge of words. Bi-directional Long-Short Term Memory wth attention is employed to determine the sentiment. We evaluate the proposed model by performing quantitative and

qualitative analysis. The experimental results show that the proposed model outperforms various word embedding models in the sentiment analysis of tweets.

Above mentioned methods can be applied to any social media classification task. The performance of proposed models is compared with different models which support the effectiveness of the proposed models and bound the information loss in their generated high-quality representations.